MAA704, Multivariate analysis.

Christopher Engström

December 20, 2013

Todays lecture

- Principal component analysis (PCA)
- Partial least squares regression (PLS-R)
- Linear discriminant analysis (LDA)

Principal component analysis

Principal component analysis (PCA) is a method often used to reduce the dimension of a large dataset to one of a more manageble size.

- The new dataset can then be used to make your analysis, supposedly with little loss of information.
- The new dataset can also be used to find hidden relationsin the data that might not have been obvious otherwise.

Principal component analysis

Assume we have n sets of data, each containing m measurements, for example:

- The value of n stocks over a period of time.
- The water usage of n households every day over some period of time.
- The rainfall at n different places over time.
- And many more.
- What if we instead could work with a smaller set of $j < n$ sets of data with m measurements each instead?

Principal component analysis

What about using the Gram-Schmidt process?

- If one set of data is a linear combination of the others, we could use the Gram-Schmidt process.
- We would then have a lower dimension dataset with no loss of information!
- Unfortunately even if the dataset contains very strong dependencies, this is generally not the case in any data collected from a real process.
- Any measurement from real data contains some random errors which quickly destroys any use we could have of the Gram-Schmidt process.

Principal component analysis

But what about if they are "nearly" linear dependent.

- For example points close to a plane in three dimensions.
- We could represent our points in $\mathbb{R}^3$ as points on the plane (in $\mathbb{R}^2$) by projecting the points on the plane?
- We could now work on this 2-dimensional space instead with (probably) a small loss of information.
- PCA works in a similar way as we will describe next.

Principal component analysis

Before doing any PCA we need to pre-process our data a bit.

- We start by subtracting the mean from every individual set.
- We then let every set of measures correspond to one row of a matrix (resulting in a $n \times m$ matrix X).

Principal component analysis

Next we want to find the single one vector that "best" describes the data in that projecting all the measures in X on the line described by this vector captures the maximum variance in the original data.

- Setting the length of the vector to equal one, this vector $\mathbf{w} = (w_1, w_2, \ldots w_m)$ should then satisfy:

$$\mathbf{w} = \underset{||w||=1}{\arg\max} \left\{ \sum_{i=1}^{n} (x_i \cdot w)^2 \right\}$$

- Writing this in matrix form yields:

$$\mathbf{w} = \underset{||w||=1}{\arg\max} \left\{ ||\mathbf{X}\mathbf{w}||^2 \right\} = \underset{||w||=1}{\arg\max} \left\{ \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} \right\}$$

Principal component analysis

Since $||\mathbf{w}|| = 1$ we can write this as:

$$\mathbf{w} = \underset{||w||=1}{\arg\max} \left\{ \frac{\mathbf{w}^\top \mathrm{X}^\top \mathrm{X} \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \right\}$$

- This is called the Rayleigh quotient for $\mathrm{M} = \mathrm{X}^\top \mathrm{X}$ if M is a Hermitian matrix ($\mathrm{M}^H = \mathrm{M}$).
- In practice M will always be a Hermitian matrix for any real measured data.
- But how do we find the vector which gives the maximum?

Principal component analysis

We let

$$R(\mathrm{M}, \mathbf{w}) = \frac{\mathbf{w}^\top \mathrm{X}^\top \mathrm{X}\mathbf{w}}{\mathbf{w}^\top \mathbf{w}}$$

We write $\mathbf{w} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_m\mathbf{v}_m$ as a linear cmbination of the eigenvectors of M resulting in:

$$R(\mathrm{M}, \mathbf{w}) = \frac{(\sum_{j=1}^{m} c_i\mathbf{v}_i)^\top \mathrm{X}^\top \mathrm{X}(\sum_{j=1}^{m} c_i\mathbf{v}_i)}{(\sum_{j=1}^{m} c_i\mathbf{v}_i)^\top (\sum_{j=1}^{m} c_i\mathbf{v}_i)}$$

Since the vectors $\mathbf{v}_i$ are orthogonal and using that $\mathrm{A}\mathbf{v}_i = \lambda_i\mathbf{v}_i$ we can write this as:

$$= \frac{\sum_{j=1}^{m} c_i^2\lambda_i}{\sum_{j=1}^{m} c_i^2}$$

Principal component analysis

Our problem can then be written as:

$$\mathbf{w} = \underset{\sum_{i=1}^{m} c_i^2 = 1}{\arg\max} \left\{ \frac{\sum_{j=1}^{m} c_i^2 \lambda_i}{\sum_{j=1}^{m} c_i^2} \right\}$$

This is clearly when $c_i = 1$ for the largest eigenvalue λ_i which means that $\mathbf{w}$ should be equal to the eigenvector corresponding to the largest eigenvalue of M.

Principal component analysis

So finding the eigenvalues and eigenvectors of $M = X^\top X$ we can find the first "principal component" as the eigenvector to the largest eigenvalue.

- Similarly the direction which gives the larges variance orthogonal to the first vector is the eigenvector corresponding to the second largest eigenvalue.
- $M = X^\top X$ is called the Covariance matrix for X.
- Now that we have our principal components, we can reduce the dimension of our data with a (hopefully) low loss of information.

Principal component analysis

To reduce the dimension we create a new basis with only the vectors pointing in directions of large variance.

- If we wanted to use all the eigenvectors we would get:

$$\mathrm{T} = \mathrm{XW}$$

  where W contains all the eigenvectors of M.

- By only taking the first L eigenvectors we get:

$$\mathrm{T}_L = \mathrm{XW}_L$$

  where W_L is a $L \times m$ matrix taken as the first L columns of W

Principal component analysis

But how do we know how many eigenvectors to include?

- The eigenvalues gives and estimate on how much of the "variance" is captured by corresponding eigenvector, we can use this!
- You might want to keep at least some proportion of the "variance":

$$\frac{V(L)}{V(n)} \geq 0.9$$

  where $V(i)$ is the sum of the largest i eigenvalues.
- Another method is to plot V and only include the eigenvectors corresponding to eigenvalues that give a significant increase of V.

Principal component analysis

- In practice you often use the correlation matrix rather than covariance matrix, this fixes problems with some measurements having much higher variance than others.
- Briefly described here was the covariance method, there is another based on the singular value decomposition (SVD) which is more stable.
- While PCA can be useful to reduce the dimension of many datasets, it doesn't always work, we need that there is at least some correlation between the different measures for it to be useful.

Principal component analysis

Short overview of the Covariance method.

- 1) Remove the mean from every individual set of measurements, order the measurements as the rows of matrix X.
- 2) Calculate the Covariance matrix $M = X^\top X$, (or use the correlation matrix).
- 3) Calculate the eigenvalues and eigenvectors of M.
- 4) Choose amount of eigenvectors to include, for example enough to at least keep 0.9 of the "variance".
- 5) Calculate $T_L = XW_L$ which is our new dataset.

Partial least squares regression

Partial least squares regression (PLS-R) is a method used to predict a set of variables (observations) from another set of variables (predictors).

- Y is a $I \times K$ matrix with I observations of K dependent variables.
- X is a $I \times J$ matrix with J predictions for the I observations.
- The goal s to predict the K dependent variables using the J predictor variables.
- When the number of predictors J is large compared to the number of observations I, common multiple regression often fails because X tend to be singular.

Partial least squares regression

The goal is to find a T such that:

$$X = TP^\top + E$$

where $TP^\top$ is a projection of X on the basis in T and E is an error term.

We then estimate Y using:

$$\hat{Y} = TBQ^\top$$

We will look at how to find the different matrices shortly.

Partial least squares regression

Overview of similar methods and their differences.

- **Principal component regression:** Use PCA on the predictors X. This gives informative directions for our predictors, but these directions might not explain the predicted variables well. Based on the spectral factorization of $X^\top X$.
- **Maximum Redundancy Analysis** Use PCA on the predicted variables Y. Seeks directions in X which well explains the responses in Y well, but we might not get an accurate prediction. Based on the spectral factorization of $Y^\top Y$.
- **Partial least squares regression** We choose our vectors in T such that the covariance between X and Y is maximized. Based on the Singular value decomposition (SVD) of $X^\top Y$.

Partial least squares regression

What do we mean by maximizing the covariance between X and Y and how do we find such vectors?

- In order to maximize the variance we want to find vectors such that:

$$\mathbf{t} = \mathrm{X}\mathbf{w},\ \mathbf{w}^\top\mathbf{w} = 1$$

$$\mathbf{u} = \mathrm{Y}\mathbf{q},\ \mathbf{q}^\top\mathbf{q} = 1$$

  Such that $\mathbf{t}^\top\mathbf{u}$ is maximized.
- $\mathbf{w}$ can be shown to be equal to the first right singular vector to $\mathrm{X}^\top\mathrm{Y}$.
- $\mathbf{q}$ can be shown to be equal to the first left singular vector to $\mathrm{X}^\top\mathrm{Y}$.

Partial least squares regression

To predict Y we wanted to use:

$$\hat{Y} = TBQ^{\top}$$

- T is a matrix containing all the vectors $\mathbf{t}_i$ as it's columns.
- B is a diagonal matrix with elements $b_i = \mathbf{t}_i^{\top}\mathbf{u}_i$ used to predict Y from $\mathbf{t}_i$
- Q is a matrix containing all the vectors $\mathbf{q}_i$ as it's columns.

Partial least squares regression

- In practice all singular values are not computed, instead a couple are computed iteratively and corresponding vectors are "deflated" from the matrices X, Y.
- This can be done in a number of different ways, which we however will not look at here.

Partial least squares regression

Similarly to PCA, we can say something about the total "variance" of X, Y explained by the first L vectors using:

- $$V_Y(L) = \frac{\sum_{i=1}^{L} b_i^2}{SS_Y}$$

  where SS_Y is the sum of the squares of measurements in Y after subtracting the mean and dividing by the standard deviation. $V_Y(L)$ is then the proportion of "variance" of Y explained by the first L vectors.

Partial least squares regression

- $$V_X(L) = \frac{\sum_{i=1}^{L} \mathbf{p}_i^\top p_i}{SS_X}$$

  where SS_X is the sum of the squares of measurements in X after subtracting the mean and dividing by the standard deviation. $V_X(L)$ is then the proportion of "variance" of X explained by the first L vectors.

- $\mathbf{p}_i = \mathrm{E}_i^\top \mathbf{t}_i$, where $\mathrm{E}_i = \mathrm{X} - \sum_{j=1}^{i-1} \mathbf{t}_j \mathbf{p}_j^\top$, (X normalized as above).

Linear discriminant analysis

Sometimes you might want to cluster your data, but you already know your clusters beforehand, you just want to be able to predict to which cluster a given measure belongs to.

- This is called a classification problem, where we typically have some observations and which class they belong to. We use this to "train" our method, which we can then use to classify new observations where wo don't know which class they belong to.
- Similar to clustering of data, but we already know our clusters and their interpretation.
- We will look at Linear Discriminant Analysis (LDA) which is used to classify observations in one of 2 different classes.
- Can quite easily be extended for multiple classes as well.

Linear discriminant analysis

Our object is to reduce the dimension while preserving as much of the information discriminating the classes from each other as possible.

- Assuming we have N D-dimensional samples.
- N_1 of these samples belonging to class c_1 and N_2 belonging to class c_2.
- Our aim is to project the samples $\mathbf{x}$ on a line $y = \mathbf{w}^\top \mathbf{x}$. Resulting in a scalar value for every sample.
- We want to find the line which when projected upon, best seperates the two classes.

Linear discriminant analysis

To evaluate the class separability we will use something called the Fisher linear discriminant:

- Maximizing this measure will give us the "best" line.
- But first we need to look at some components we will need.

Linear discriminant analysis

We start by looking at the distance between the projected means of the classes.

- In the original space the mean $\boldsymbol{\mu}_i$ of a class is easily found:

$$\boldsymbol{\mu}_i = \frac{1}{N_i} \sum_{x \in c_i} \mathbf{x}$$

- And for the projected means we get:

$$\tilde{\mu}_i = \frac{1}{N_i} \sum_{y \in c_i} \mathbf{y} = \frac{1}{N_i} \sum_{x \in c_i} \mathbf{w}^\top \mathbf{x} = \mathbf{w}^\top \boldsymbol{\mu}_i$$

- We could now use the distance between the projected means:

$$|\tilde{\mu}_1 - \tilde{\mu}_2| = |\mathbf{w}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)|$$

Linear discriminant analysis

Although the distance between the projected means might seperate the classes well, it does not take into consideration the variance of the data.

- If there is a high variance in the same direction as the one we would get when maximizing the direction of the means, we could get bad seperability anyway.
- To solve this we will look at the variance within a class (also called the scatter) on the projected line:

$$\tilde{S}_i^2 = \sum_{y \in c_i} (y - \tilde{\mu}_i)^2$$

- Adding the scatter of both classes and we get the "within-class scatter":

$$(\tilde{S}_1^2 + \tilde{S}_2^2)$$

Linear discriminant analysis

Fishers linear discriminant is defined as the function $\mathbf{w}^\top \mathbf{x}$ which maximizes $J(\mathbf{w})$:

$$J(\mathbf{w}) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{S}_1^2 + \tilde{S}_2^2}$$

We are now looking for a projection where elements from the same class are projected close to eachother (low within-class scatter) and the distance between the projected class means are far apart.

Linear discriminant analysis

To maximize $J(\mathbf{w})$ we start by writing it using $\mathbf{w}$.

- We start by looking at the "within-class scatter":

$$\tilde{S}_i^2 = \sum_{y \in c_i} (y - \tilde{\mu}_i)^2 = \sum_{y \in c_i} (\mathbf{w}^\top \mathbf{x} - \mathbf{w}^\top \boldsymbol{\mu}_i)^2$$

-

$$= \sum_{y \in c_i} \mathbf{w}^\top (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^\top \mathbf{w}$$

- We call: $\mathrm{S}_i = \sum_{x \in c_i} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^\top$ and $\mathrm{S}_W = \mathrm{S}_1 + \mathrm{S}_2$

  and get:

$$\tilde{S}_1^2 + \tilde{S}_2^2 = \mathbf{w}^\top (\mathrm{S}_1 + \mathrm{S}_2)\mathbf{w} = \mathbf{w}^\top \mathrm{S}_W \mathbf{w}$$

Linear discriminant analysis

If we instead look at the distance between projected means we get:

$$|\tilde{\mu}_1 - \tilde{\mu}_2|^2 = (\mathbf{w}^\top \boldsymbol{\mu}_1 - \mathbf{w}^\top \boldsymbol{\mu}_2)^2$$

$$\mathbf{w}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \mathbf{w}$$

We call $\mathsf{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top$ and we get:

$$|\tilde{\mu}_1 - \tilde{\mu}_2|^2 = \mathbf{w}^\top \mathsf{S}_B \mathbf{w}$$

Linear discriminant analysis

Combining the two gives the following function in $\mathbf{w}$ which we want to maximize:

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_W \mathbf{w}}$$

For which we can find the maximum by setting the gradient to zero:

$$\frac{d}{d\mathbf{w}} J(\mathbf{w}) = 0$$

Linear discriminant analysis

With the help of some matrix calculus this is equivalent to the eigenvalue problem:

$$\mathsf{S}_W^{-1}\mathsf{S}_B\mathbf{w} = J\mathbf{w}$$

With solution:

$$\mathbf{w}^* = \arg\max\left[\frac{\mathbf{w}^\top \mathsf{S}_B\mathbf{w}}{\mathbf{w}^\top \mathsf{S}_W\mathbf{w}}\right] = \mathsf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

Linear discriminant analysis

A short note on generalizations to more than two classes.

- Fisher's linear discriminant itself can be generalized for more classes, will result in finding a subspace $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{C-1}$ where C is the number of classes instead.
- Another alternative is to classify every class itself with respect to all the other classes, giving C classifiers that can then be combined.
- It is also possible to use the classifiers of every pair of classes.

Linear discriminant analysis

While LDA works well for Gaussian distributions, some other shapes or types of distributions can give problems.

- Some shapes of the classes can lead to problems, for example two interwined "C" shapes.
- If both clusters have the same mean we cannot classify them.
- Or if the information lies not in the mean of the classes but in the variance it will also fail.